

School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA 30332, USA
{jiyfeng, jacobec}@gatech.edu

Discourse relations bind smaller linguistic elements into coherent texts. However, automatically identifying discourse relations is difficult, because it requires understanding the semantics of the linked sentences. A more subtle challenge is that it is not enough to represent the meaning of each sentence of a discourse relation, because the relation may depend on links between lower-level elements, such as entity mentions. Our solution computes distributional meaning representations by composition up the syntactic parse tree. A key difference from previous work on compositional distributional semantics is that we also compute representations for entity mentions, using a novel downward compositional pass. Discourse relations are predicted not only from the distributional representations of the sentences, but also of their coreferent entity mentions. The resulting system obtains substantial improvements over the previous state-of-the-art in predicting implicit discourse relations in the Penn Discourse Treebank.

We further argue that purely vector-based representations on sentences are insufficiently expressive to capture discourse relations. To see why, consider what happens in Example (2), where a tiny change is made based on Example (1). After changing the subject of the second sentence to Bob, the original discourse relation seems no longer holding in Example (2). But despite the radical difference in meaning, the distributional representation of the second sentence will be almost unchanged: the syntactic structure remains identical, and the words “*he*” and “*she*” have very similar word representations. We address this issue by computing vector representations not only for each sentence, but also for each coreferent entity mention within the sentences. These representations are

meant to capture the *role* played by the entity in the text. We compute entity-role representations using a novel feed-forward compositional model, which combines *upward* and *downward* passes through the syntactic structure. Representations for these coreferent mentions are then combined into a classification model, and help to predict the implicit discourse relation. In combination, our approach achieves a 3% improvement in accuracy over the best previous work (Lin et al., 2009) on the second-level discourse relation identification in the PDTB.¹

Our model requires a syntactic parse tree, which is produced automatically from the Stanford CoreNLP parser Klein & Manning (2003). A reviewer asked whether it might be better to employ a left-to-right recurrent neural network, which would obviate the need for this language-specific resource. While it would clearly be preferable to avoid the use of language-specific resources whenever possible, we think this approach is unlikely to succeed in this case. A key difference between language and other types of data is that language has inherent recursive structure. A rich literature in both linguistics and natural language processing elaborates on the close relationship between (recursively-structured) syntax and semantics. Therefore, we see strong theoretical evidence — as well as practical evidence from the history of natural language processing — that syntactic parse structures are central to capturing the meaning in text.

Regarding the multilingual question, there are now accurate parsers and annotated treebanks for dozens of languages,² and training accurate parsers for “low resource” languages is a hot research topic, with substantial interest from both industry and academia. Languages differ substantially in the importance of word ordering, with English emphasizing word order more than most other languages (Bender, 2013). To our knowledge, it is an open question as to whether left-to-right recurrent neural networks will successfully extract meaning in languages where word order is more free.

2 ENTITY AUGMENTED DISTRIBUTIONAL SEMANTICS FOR RELATION IDENTIFICATION

We briefly describe our approach to entity-augmented distributional semantics and to discourse relation identification. Our relation identification model is named as DISCO2, since it is a **d**istributional **c**ompositional approach to **d**iscourse relations.

2.1 ENTITY AUGMENTED DISTRIBUTIONAL SEMANTICS

The entity-augmented distributional semantics includes two passes in composition procedure: the upward pass for distributional representation of sentence, while the downward pass for distributional representation of entities shared between sentences.

Upward pass Distributional representations for sentences are computed in a feed-forward *upward* pass: each non-terminal in the binarized syntactic parse tree has a K -dimensional distributional representation that is computed from the distributional representations of its children, bottoming out in representations of individual words. We follow the Recursive Neural Network (RNN) model proposed by Socher et al. (2011). Specifically, for a given parent node i , we denote the left child as $\ell(i)$, and the right child as $r(i)$. We compose their representations to obtain, $\mathbf{u}_i = \tanh(\mathbf{U}[\mathbf{u}_{\ell(i)}; \mathbf{u}_{r(i)}])$, where $\tanh(\cdot)$ is the element-wise hyperbolic tangent function, and $\mathbf{U} \in \mathbb{R}^{K \times 2K}$ is the upward composition matrix. We apply this compositional procedure from the bottom up, ultimately obtaining the sentence-level representation \mathbf{u}_0 .

Downward pass As seen in the contrast between Examples (1) and (2), a model that uses a single vector representation for each sentence would find little to distinguish between “*she was hungry*” and “*he was hungry*”. It would therefore almost certainly fail to identify the correct discourse relation for at least one of these cases, which requires tracking the roles played by the entities that are coreferent in each pair of sentences. To address this issue, we augment the representation of each sentence with additional vectors, representing the semantics of the role played by each coreferent entity in each

¹For more details, please refer to the long version of this paper (Ji & Eisenstein, 2015)

²<http://universaldependencies.github.io/docs/#language-other>

Model	+Entity semantics	+Surface features	K	Accuracy(%)
<i>Prior work</i>				
1. Lin et al. (2009)		Yes		40.2
<i>Our work</i>				
2. Surface feature model		Yes		39.69
3. DISCO2	No	No	50	36.98
4. DISCO2	Yes	No	50	37.63
5. DISCO2	No	Yes	50	42.53
6. DISCO2	Yes	Yes	50	43.56*

* significantly better than Lin et al. (2009) with $p < 0.05$

Table 1: Experimental results on multiclass classification of second-level discourse relations. The results of Lin et al. (2009) are shown in line 1; the results for our reimplement of this system are shown in line 2.

sentence. Rather than represent this information in a logical form — which would require robust parsing to a logical representation — we represent it through additional distributional vectors. The role of a constituent i can be viewed as a combination of information from two neighboring nodes in the parse tree: its parent $\rho(i)$, and its sibling $s(i)$. We can make a downward pass, computing the downward vector \mathbf{d}_i from the downward vector of the parent $\mathbf{d}_{\rho(i)}$, and the **upward** vector of the sibling $\mathbf{u}_{s(i)}$: $\mathbf{d}_i = \tanh(\mathbf{V}[\mathbf{d}_{\rho(i)}; \mathbf{u}_{s(i)}])$, where $\mathbf{V} \in \mathbb{R}^{K \times 2K}$ is the downward composition matrix. The base case of this recursive procedure occurs at the root of the parse tree, which is set equal to the upward representation, $\mathbf{d}_0 \triangleq \mathbf{u}_0$.

2.2 RELATION IDENTIFICATION MODEL

To predict the discourse relation between an sentence pair (m, n) , the decision function is a sum of bilinear products,

$$\psi(y) = (\mathbf{u}_0^{(m)})^\top \mathbf{A}_y \mathbf{u}_0^{(n)} + \sum_{i,j \in \mathcal{A}(m,n)} (\mathbf{d}_i^{(m)})^\top \mathbf{B}_y \mathbf{d}_j^{(n)} + \beta_y^\top \phi_{(m,n)} + b_y, \quad (1)$$

where the predicted relation is given by $\hat{y} = \arg \max_{y \in \mathcal{Y}} \psi(y)$, and $\mathbf{A}_y, \mathbf{B}_y \in \mathbb{R}^{K \times K}$ are the classification parameters for relation y . A scalar b_y is used as the bias term for relation y , and $\mathcal{A}(m, n)$ is the set of coreferent entity mentions shared among the sentence pair (m, n) . For the cases where there are no coreferent entity mentions between two sentences, $\mathcal{A}(m, n) = \emptyset$, the classification model considers only the upward vectors at the root. We also use the *surface features* vector $\phi_{(m,n)}$ in the decision function, as we find that, this approach outperforms prior work on the classification of implicit discourse relations in the PDTB, when combined with a small number of surface features.

3 EXPERIMENTS

We evaluate our approach on the implicit discourse relation identification in the Penn Discourse Treebank (PDTB). PDTB relations may be *explicit*, meaning that they are signaled by discourse connectives (e.g., *because*); alternatively, they may be *implicit*, meaning that the connective is absent. We focus on the more challenging problem of classifying implicit discourse relations. Aiming to build a discourse parser in future, we follow the same experimental setting proposed by Lin et al. (2009), and evaluate our relation identification model on the *second-level* relation types.

We run the Stanford parser (Klein & Manning, 2003) and the Berkeley coreference system (Durrett & Klein, 2013) to obtain syntactic trees and coreference results respectively. In the PDTB, each discourse relation is annotated between two argument spans. For non-sentence argument span, we identify the syntactic subtrees with the span, and construct a right-branching superstructure to unify them into a tree.

Table 1 presents results for multiclass identification of second-level PDTB relations. As shown in lines 5 and 6, DISCO2 outperforms the prior state-of-the-art (line 1). The strongest performance is

obtained by including the entity distributional semantics, with a 3.4% improvement over the accuracy reported by Lin et al. (2009) ($p < .05$). The improvement over our reimplementation of this work (line 2) is even greater, which shows how the distributional representation provides additional value over the surface features. The contribution of entity semantics is shown in Table 1 by the accuracy differences between lines 3 and 4, and between lines 5 and 6.

4 CONCLUSION

Discourse relations are determined by the meaning of their arguments, and progress on discourse parsing therefore requires computing representations of the argument semantics. We present a compositional method for inducing distributed representations not only of discourse arguments, but also of the entities that thread through the discourse. By jointly learning the relation classification weights and the compositional operators, this approach outperforms prior work based on hand-engineered surface features. More discussion and experimental results can be found in a forthcoming journal paper (Ji & Eisenstein, 2015).

REFERENCES

- Baroni, Marco, Bernardi, Raffaella, and Zamparelli, Roberto. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technologies*, 2014.
- Bender, Emily M. *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*, volume 6 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers, June 2013. doi: 10.2200/s00493ed1v01y201303hlt020. URL <http://dx.doi.org/10.2200/s00493ed1v01y201303hlt020>.
- Durrett, Greg and Klein, Dan. Easy Victories and Uphill Battles in Coreference Resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, October 2013. Association for Computational Linguistics.
- Ji, Yangfeng and Eisenstein, Jacob. One vector is not enough: Entity-augmented distributional semantics for discourse relations. *Conditionally accepted to Transactions of the Association for Computational Linguistics (TACL)*, 2015.
- Klein, Dan and Manning, Christopher D. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pp. 423–430. Association for Computational Linguistics, 2003.
- Knott, Alistair. *A data-driven methodology for motivating a set of coherence relations*. PhD thesis, The University of Edinburgh, 1996.
- Lin, Ziheng, Kan, Min-Yen, and Ng, Hwee Tou. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pp. 343–351. Association for Computational Linguistics, 2009.
- Lin, Ziheng, Ng, Hwee Tou, and Kan, Min-Yen. Automatically Evaluating Text Coherence Using Discourse Relations. In *Proceedings of ACL*, pp. 997–1006, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.
- Louis, Annie, Joshi, Aravind, and Nenkova, Ani. Discourse indicators for content selection in summarization. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 147–156. Association for Computational Linguistics, 2010.
- Mann, William. Discourse structures for text generation. In *Proceedings of the 10th International Conference on Computational Linguistics and 22nd annual meeting on Association for Computational Linguistics*, pp. 367–375. Association for Computational Linguistics, 1984.
- Prasad, Rashmi, Dinesh, Nikhil, Lee, Alan, Miltsakaki, Eleni, Robaldo, Livio, Joshi, Aravind, and Webber, Bonnie. The Penn Discourse Treebank 2.0. In *LREC*, 2008.

- Socher, Richard, Lin, Cliff C, Manning, Chris, and Ng, Andrew Y. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 129–136, 2011.
- Socher, Richard, Perelygin, Alex, Wu, Jean Y, Chuang, Jason, Manning, Christopher D, Ng, Andrew Y, and Potts, Christopher. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, 2013.
- Somasundaran, Swapna, Namata, Galileo, Wiebe, Janyce, and Getoor, Lise. Supervised and unsupervised methods in employing discourse relations for improving opinion polarity classification. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 170–179. Association for Computational Linguistics, 2009.
- Turney, Peter D, Pantel, Patrick, et al. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188, 2010.
- Webber, Bonnie, Knott, Alistair, Stone, Matthew, and Joshi, Aravind. Discourse relations: A structural and presuppositional account using lexicalised tag. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 41–48, 1999.
- Yoshida, Yasuhisa, Suzuki, Jun, Hirao, Tsutomu, and Nagata, Masaaki. Dependency-based Discourse Parser for Single-Document Summarization. In *EMNLP*, 2014.